Choice-Based Conjoint for Designing Home Appliances: Human Versus Silicon Samples of Respondents

By Daniel Baier*, Danilo Randazzo, and Maximilian Unger

Received March 6, 2026/Accepted June 3, 2026 (after two revisions)

Some researchers assume that Generative Artificial Intelligence (GenAI) has the potential to augment or even replace traditional market research. In this paper, we explore whether this assumption holds for Choice-Based Conjoint Analysis (CBC) studies, taking an investigation in the home appliance industry as an example: Choices among sets of multi-attributed stimuli are generated for digital twins of a sample of potential home appliance buyers (a so-called silicon sample of respondents). Then, partworth estimates from these choices are compared to estimates from corresponding human choices. As an experimental factor, GenAI generates these choices under varying amounts of provided information from company experts and preliminary studies. The results are encouraging: Even for a CBC study with innovative attribute-levels, partworth estimates from GenAI choices come close to partworth estimates from human choices, at least at the aggregate level and when market-related information of experts and/or from preliminary studies is provided. However, at the individual level and when additional information is not provided, GenAI choices lead to partworth estimates whose validity is almost equivalent to that of partworth estimates from random choices.

1. Introduction

The introduction of the Transformer architecture for neural networks by a team of researchers at Google (Vaswani et al., 2017) as well as the related formation of Large Language Models (LLMs) have significantly promoted the development and spread of artificial intelligence in both private and professional settings (see, e.g., Chang et al., 2024): LLMs like Generative Pre-trained Transformer (GPT) from OpenAI or Gemini from Google together with derived chatbots (e.g. ChatGPT or Bart/Gemini) are convincing in completing so-called Generative Artificial Intelligence (GenAI) tasks like text summarization, question answering, text or image generation, or drawing conclusions from presented text. So, Chang et al. (2024) state that LLMs and derived chatbots nowadays are able – even when compared to human experts – to respond fluently and meaningful to questions with correct content and language. However, they still seem inferior to human

Daniel Baier is Professor of Marketing & Innovation at the University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany, Phone: +49/921 55-4340, E-Mail: daniel.baier@uni-bayreuth.de.
* Corresponding author.

Danilo Randazzo is responsible for Product Innovation at Liebherr-Hausgeräte Ochsenhausen GmbH, Memminger Straße 77–79, 88416 Ochsenhausen, Germany, E-Mail: danilo.randazzo@liebherr.com. He is also Doctoral Candidate at the Chair of Marketing & Innovation, University of Bayreuth, Germany.

Maximilian Unger is Research Assistant and Doctoral Candidate at the Chair of Marketing & Innovation at the University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany, E-Mail: maximilian.unger@uni-bayreuth.de.

experts when it comes to capturing contradictions and semantic subtleties in texts or images.

In marketing and market research, GenAI's support potential was recognized very early and used to analyze customers' feedback to products and services (e.g., Bouschery et al., 2023; Jeong & Lee, 2024), to answer social media posts (e.g., in Koc et al., 2023), to replace respondents in surveys (see, e.g., Arora et al., 2025, Brand et al. 2023; Guo et al., 2023), to generate web content (Sundberg & Holmström, 2024; Tafesse & Wood, 2024), or even to develop new ideas and concepts for products and services (e.g., Bouschery et al., 2023; Filippi, 2023).

However, some authors also question the validity of using LLMs in marketing and market research, especially when the task is to replace samples of human respondents in surveys (e.g., Grewal et al., 2025; Park et al., 2024; Sarstedt et al., 2024; Viglia et al., 2024; Wang et al., 2024). So, e.g., Park et al. (2024) show in their comparison of answers by human respondents in 14 published surveys and – alternatively – generated answers by GPT-3.5 that only 37.5 % of the findings of the 14 published studies could be replicated. Park et al. (2024) as well as Sarstedt et al. (2024) therefore call for further research on this topic, especially with newer and advanced LLMs.

In this paper, we contribute to this call by exploring whether GPT answers to choice tasks in a market research study lead to similar partworth estimates as if estimated from human answers. Our application area for this comparison is the market-oriented design of a new luxurious refrigerator for German households by a major European home appliance manufacturer. We selected this product and target segment since – on one side – buying an expensive refrigerator typically is a rare purchase task for a household (a typical refrigerator has a lifespan of more than 10 years) that cannot be easily predicted from past behavior (being therefore an advanced task for an LLM) whereas – on the other side – help is available since producers' knowledge about customer preferences in this market is quite good, steaming, e.g., from past market research and customer complaints as well as published consumer-oriented comparisons of products by consumer organizations (e.g., Stiftung Warentest or Verbraucherzentrale in Germany).

We generate GPT choices among sets of multi-attributed home appliances by simulating a representative sample of potential buyers (the so-called silicon sample) and – for comparisons – collect similar choices from a sample of human respondents in advance with the same distribution as the silicon sample regarding age, gender, household size, income, and innovativeness. Based on both samples and their generated/collected choices, hierarchical Bayes multinomial logit partworths are estimated and compared. As an experimental factor, GPT generates choices under varying amounts of information provided, i.e., "without any additional information", "with market-releated information from company experts", and "with information from a preliminary study".

The paper is structured as follows: In two sections (Section 2 and 3), we discuss the current state-of-the-art of conjoint analysis, especially for home appliances, as well as the opportunities and threats to collect market research data from silicon samples instead of human samples. Section 4 discusses data collection from human and silicon samples in our application. Section 5 discusses the results including estimation of partworths, comparisons, and potential limitations. Section 6 presents an outlook and conclusions.

2. Background: Conjoint analysis for home appliances

Since its introduction to marketing in the early 1970s (Green & Rao, 1971; Green & Wind, 1975), conjoint analysis has become a widely used market research tool. So, e.g., Orme (2020, p. 151) extrapolates from a survey among market researchers that more than 27,000 conjoint analysis studies are conducted worldwide per year representing an attractive market segment for market research institutes. Moreover, Roberts et al. (2014) conclude from their citation analysis and surveys among marketing managers, intermediaries, and academics, that conjoint analysis articles and tools had and have highest impact on marketing practice. Typical application fields with proven helpfulness for the marketers have been – for a long time – product design, pricing research, new product development, or package design (Rao, 2026, pp. 8 ff.).

The basic idea is to estimate preference impacts of attribute-levels (so-called partworths) from collected preferences and/or choices among hypothetical products (so-called stimuli or options). Five major steps define a typical conjoint analysis study (Green & Srinivasan, 1978; Green & Srinivasan, 1990): (1) Assumption of an attribute-level based preference and/or choice model for products in a market of interest, (2) construction of attribute-level-combinations as stimuli, (3) collection of preferences and/or choices among these stimuli from a sample of potential buyers, (4) estimation of preference impacts of the attribute-levels on overall preference and/or choice, and (5) market simulation based on the estimated partworths and the assumed preference and/or choice model.

Being a major market research tool over time, the methodology received many published discussions, adaptations, and improvements over time resulting in a confusing galaxy of variants with respect to preference and/or choice models as well as stimulus generation, data collection, and parameter estimation methods (see Steiner & Meißner, 2018, Baier & Brusch, 2021, or Rao, 2026 for reviews). However, since more than 25 years, the Choice-Based Conjoint Analysis (CBC) variant of conjoint analysis has been the most popular and most often used one (see, e.g., Rao, 2026, pp. 167 ff.). So, e.g., in a recent customer survey, Sawtooth Software, a major con-

joint analysis software provider, concludes from a user survey that in 74 % of all conjoint analysis studies, CBC is the variant applied (Sawtooth Software, 2025).

CBC is based on discrete choice analysis and design of experiments from multivariate statistics and was originally proposed for market research by Louviere & Woodworth (1983). Respondents are repeatedly confronted with choice tasks that consist of a fixed number of stimuli and asked to select the mostly preferred one. The observed choices across all respondents then are analyzed using a multinomial logit model and maximum likelihood estimation to derive partworths at an aggregate level. Later (Allenby et al., 1995; Sawtooth Software, 2021), hierarchical Bayes estimation was introduced as a major CBC improvement, allowing to model heterogeneity among respondents and to estimate partworths at an individual level.

In many consumer goods and service markets, CBC has been successfully applied. So, e.g., Selka & Baier (2014) analyzed 1,777 commercial CBC studies, performed by a major European market research institute and found that 53.0 % of these studies dealt with consumer goods, 40.9 % with consumer services. Most often researched markets were food and beverage goods (37.5 % of the consumer goods studies), cars and other vehicles (23.7 %), drugstore and pharmacy goods (12.8 %), household and hygiene goods (12.4 %), as well as home appliances like freezers, ovens, microwaves, refrigerators, and washing machines (8.2 %). Application fields were pricing research (88 %, multiple assignments allowed), product design (59 %), market segmentation (12 %), advertisement research (7 %), and others. 77 % made use of computer-aided web interviewing (in most cases based on Sawtooth Software's Lighthouse system), in 68.3 % of the studies, the questionnaires were distributed with help of online access panels.

Published articles on home appliance CBC studies are rare (see Table 1 for an overview). So, e.g., Razali et al. (2022) discuss a CBC study where refrigerator purchase decisions of households in Malaysia were under investigation based on the four attributes energy label (awarded by NGOs), energy consumption, energy saving features, and price. n=426 respondents were asked to select the most preferred attribute-level-combination in four choice sets consisting of three stimuli. A multinomial logit analysis of the collected data showed that price and energy saving features were the attributes that most influenced the choice of a refrigerator. Zha et al. (2020) conducted two similar CBC studies in China for refrigerators and washing machines with the five attributes brand, capaci-

Industry	Task	Method (Sample/Data)	Results	Reference
Refrigerators, washers, TV	Analysis of brand switching behavior and primary sources	Crosstable analysis of current and last purchased brand (USA)	Brands differ w.r.t. their customer source (switchers, loyals)	Bayus (1992)
Refrigerators	Preference estimation based on brand, price, energy label, volume, defrost	ACA (n= 249, Germany), CCA and CCC (n=242, Germany)	CCC outperforms ACA and CCA w.r.t. prediction of market shares	Hensel-Börner & Sattler (2000)
Refrigerators	Price-premium estimation for highest energy efficiency labels (A+)	Realized sales price prediction from labels, volume, defrost (Spain)	Price-premium is ~9% of the average price or about 60 Euro	Galarraga et al. (2011)
Washing machines	Preference estimation based on brand, price, label, volume	TCA (n=118, India)	Price and label (power consumption) were most important	Kulshreshtha et al. (2019)
Refrigerators	Preference estimation based on price, label, volume. lifetime, service reliability, defrost	TCA (n=198, Turkey)	Label, lifetime, and service reliability were most important	Aydin & Seker (2020)
Refrigerators	Preference estimation based on price and labels (stars, savings)	CBC (n=429, Malaysia)	Energy savings were most important	Razali et al. (2022)
White goods	Attribute importance when buying home appliances	MaxDiff (n=9,050, France, Germany, Italy, Japan, Poland, Sweden, Turkey, UK, USA)	Durability, energy, water, volume, noise were most important	Herring et al. (2025)
Refrigerators (multi-door, interior)	Preference estimation based on accessories, water and ice, door opening	CBC (n=617, human and comparable silicon sample/ digital twins, Germany)	IceCenter and flexible storage box are important, silicon samples provide similar results	This paper

Abbreviations: ACA=Adaptive Conjoint Analysis, CBC = Choice Based Conjoint Analysis, CCA = Customized Conjoint Analysis, CCC = Customized Computerized Conjoint Analysis, MaxDiff = Maximm Difference Scaling = Best Worst Scaling, TCA = Traditional Conjoint Analysis, w.r.t.=with respect to

Tab. 1: Selected published market research on CBC applications for home appliances.

ty, energy label (a mandatory system that grades home appliances with respect to energy consumption across different scenarios), power consumption (during normal operation), and price. n=453 respondents were confronted with four choice sets for refrigerators and four for washing machines. Each choice set consisted of four stimuli. The multinomial logit analysis of these data showed that for the respondents' choices among refrigerators the attributes brand, energy label, and power consumption were equally important, among washing machines the attributes energy label and power consumption. In one market (Zha et al., 2020), price was also important.

Whereas CBC studies nowadays are acknowledged to be very helpful for improving home appliances as well as other goods or services, for a long time, the widespread use of online access panels in these studies for collecting data has been questioned (see, e.g., Selka & Baier, 2014): So, Göritz (2004), Göritz (2006), and Callegaro et al. (2014) discuss the low representativity of these panels (e.g., caused by using nonprobability- or probability-based recruiting and maintenance instead of invitation-only processes) and the low validity caused by incentivizing the survey participants. Arndt et al. (2022) refer – as a consequence – to the necessity of selection screening and accuracy screening when using online access panels (often neglected), especially when using (cheaper) panels without an actively-managed recruiting and maintenance process. However, recent comparisons (e.g., Rudolph & Seelkopf, 2025) have shown – at least for German online access panels and when screeners are applied – that identical conjoint studies distributed across eight online access panels lead to similar results, despite prices per respondent varying up to a factor of 3.7 across the eight panels. Moreover, when estimates are compared to other estimates, it is often referred to the problem that other data collection formats (CATI, CAPI, student samples) also have problems with representativity and validity (see, e.g., the discussion in Kato & Miura, 2021).

Consequently, some authors accept the results of a CBC study that makes proper and controlled use of an online access panel as a "ground truth" for receiving partworth estimates (see, e.g., Arora et al., 2025; Brand et al., 2023) when comparing data collection and analysis methodologies. These authors also argue, that – with a reference to cost and quality issues of an online access panel for data collection in a CBC study – recent progress in GenAI could provide an alternative solution for data collection as discussed in the next section.

3. Background: Silicon samples in market research

Transformer-based large language models (LLMs for short, such as GPT or Bard/Gemini) and derived chatbots (e.g. ChatGPT) have demonstrated that they can fulfill many natural language processing tasks in an acceptable manner, e.g., by summarizing and classifying texts, by generating useful and correct answers to questions, or by analyzing data and formatting the results in a pre-specified format (see, e.g., the overview by Chang et al., 2024). In business administration, these proven abilities could lead to important productivity gains. So, e.g., in a summary of published use cases, McKinsey & Company (2023) estimate the expected worldwide productivity gains by these tools at about US$ 2.6 to 4.4 trillion per year. More than 75 % of these gains were attributed to customer management, marketing and sales, software development, and product development. The latter application area ("product R&D" in the original report) mainly consists of activities such as product-related market research, idea generation, and concept design in the early phases of product development, but also prototyping and testing in the later phases.

Other authors, e.g., Arora et al. (2025) as well as Sarstedt et al. (2024), also conclude in their review articles that LLMs and derived chatbots can assist market research activities across all research stages of a study project. Table 2 summarizes studies with results from these applications: During study design and sample selection, LLMs can help to create the questionnaire (e.g., selection of buying relevant attributes and levels for a CBC study) and determine sample characteristics (e,g., definition of quotas that describe the target segment). Sarstedt et al. (2024) see "a considerable promise in using LLMs such as GPT in upstream parts of the research process," e.g., for scale pretesting and questionnaire design. They highlight the usefulness of consulting an LLM as a test person when pretesting scales or as an expert when assessing the appropriateness of questions.

For data collection, Arora et al. (2025) as well as Sarstedt et al. (2024) discuss in their reviews the possibilities to use LLMs to simulate synthetic respondents by answering questionnaires with pre-defined socio-demographic and psychographic characteristics. Both reviews refer to Brand et al. (2023) as the only example where an LLM was used to simulate CBC studies and where the results from silicon samples were compared to results from human samples with choice tasks collected with help of an online access panel as ground truth. Brand et al. (2023) relied on the GPT-3.5 to fill-in questionnaires with CBC tasks in four markets (toothpastes, deodorants, laptops, tablets). In each market, products were defined by three to six attributes, each with two or three levels. The prompt for each GPT-3.5 response (that simulated one "silicon" respondent) consisted of (1) the information that he/she should answer as a randomly selected potential buyer in the market under study with defined characteristics (e.g., income, gender, political affiliation, race and ethnicity, education) and (2) the instruction to select a preferred stimulus (presented as attribute-level-combinations resp. hypothetical products) in the following choice tasks. As choice tasks all possible pairs of attribute-level combinations in the market under study were presented. Additionally, in each task, GPT-

Industry	Task	Method	Results	Reference
Political assessments	Sampling characteristics of republicans & democrats	Use of GPT-3 and humans for sampling characterizing words	GPT-3 & humans deliver similar distributions of characteristics	Argyle et al. (2023)
Reward schemes	Estimate preferences for customer reward schemes	Use of GPT-3.5 & GPT-4 to collect choices among rewards	Chain-of-thought prompts deliver similar choice distributions	Goli & Singh (2024)
No specific industry	Questionnaire development and knowledge provision	Use of ChatGPT to develop scales and provide knowledge	ChatGPT provides answers that are rated as particularly helpful	Guo et al. (2023)
Dog food	Sampling attitudes & willingness-to-buy for new dog food	Use of GPT-4 for sampling in silicon vs. traditional surveys	Silicon sample delivers similar attitudes & willingness-to-buy	Arora et al. (2025)
Toothpaste, notebooks	Estimate preferences & WTP for new products	Use of GPT-3.5 to collect choices vs. a traditional survey	Silicon sample provides similar preference orders and WTP	Brand et al. (2023)
Soft drinks	Collect preferences for established and less known products	Use of GPT-4o to replicate a human sample (digital twins)	GPT-4o preferences reproduce aggregate trends	Kaiser et al. (2025)
No specific industry	Generate synthetic answers to questionnaires that pass tests	Use of a GenAI agent who answers any online questionnaire	GenAI agent produced answers that pass 99.8% of the tests	Westwood (2025)
Refrigerator	Estimate preferences via CBC data collection	Use of GPT-5.2 to replicate choices of humans (digital twins)	GPT-5.2 generated silicon samples can provide similar results	This paper

Tab. 2: Selected comparisons of human samples with silicon samples in market research.

3.5 was allowed to choose the "none" option (indicating that none of the two presented stimuli is acceptable). For comparison reasons, choice data from humans was collected via an online access panel (Prolific). The panelists were confronted with 12 choice tasks and had to indicate which stimulus they preferred or whether none was acceptable. The collected data from the human and silicon samples, then, were analyzed using the multinomial logit model with a focus on predicting the willingness-to-pay for all attribute-levels at an aggregate level. Brand et al. (2023) summarize the results of this comparison by stating that the importance of the attributes and levels was ranked correctly, especially when GPT-3.5 was supplemented by additional information in the prompt (e.g., results from past studies in related markets).

However, Brand et al. (2023) and other authors (Goli & Singh, 2024; Park et al., 2024; Sarstedt et al., 2024) mention surprising effects when using GPT to generate answers in main studies. So, the order of possible answers (stimuli) when prompting a question (choice task) influences the answer. Moreover, a so-called "correct answer" effect could be observed, even for questions where humans tend to give varying answers: The LLM responds with an assumed best answer and repeats this answer even if the context, the order of possible answers, the role as a respondent (e.g., sex, age, income, and so on), or the "temperature" (a parameter that influences the probabilities with which an LLM gives the same answer) was modified. Other, especially non-conjoint comparisons between silicon samples and human samples support these findings and "cast doubts on the validity of using LLMs as a general replacement for human participants in the social sciences" (Park et al., 2024). So, Park et al. (2024) compared the results based from human and silicon samples (generated by GPT-3.5) when replicating 14 well-known social science studies. They found that only 37.5 % of the expected effects could be confirmed by the silicon sample in contrast to 50 % confirmed by the human sample. Sarstedt et al. (2024) conclude from these comparisons and investigations that further research is needed before silicon samples can replace human samples in market research. As they only refer to older LLM versions (e.g., GPT-3.5) in their comparisons and investigations, they call for research with newer LLMs and derived chatbots.

Here, the newer comparison and investigation by Arora et al. (2025) that focuses on GPT-4, comes to more favorable conclusions in non-conjoint comparisons: They find that silicon samples generated by GPT-4 "picked the answer direction well; when the average for a variable is toward the lower end of the scale, the synthetic data average tends also to be low and vice versa" (Arora et al., 2025, p. 24). Moreover, Arora et al. (2025) were able to improve the silicon sample findings (compared to a ground truth derived from human samples) when information from past studies was made available to the LLM. However, these promising results from Arora et al. (2025) refer to qualitative and quantitative market research without CBC tasks.

The contribution of this paper, therefore, is to use a newer LLM (GPT-5.2) to generate silicon samples for a CBC study and to compare the derived results with results derived from a human sample. Moreover, in contrast to the CBC study by Brand et al. (2023) who used GPT-3.5 for generating responses to all possible choice tasks, in our study we also try to align the LLM data collection process with the data collection process from human respondents, e.g., by using typical CBC data collection formats (choice tasks with four stimuli, balanced overlap design) as well as best prompting strategies. As already discussed in the introduction, we selected home appliances for our comparison since this branch is rather often investigated by CBC studies (see Section 2) and therefore information from past studies and – if needed – preliminary studies is available.

Within our research context – based on our literature review and the established household appliance market under investigation – we anticipate that GPT-5.2 will be capable of generating valid selection decisions from presented stimuli and valid partworth estimates based on them, provided that the simulated customer is meticulously characterized beforehand (e.g., in terms of sociodemographic and psychographic characteristics, as well as past purchasing behavior and current preferences regarding products, brands, attributes, and feature specifications).

4. Data collection

4.1. Market under study and preliminary study

The term home appliance refers to large household electrical appliances, such as refrigerators, washers, dishwashers, freezers, or fridge-freezers, sometimes also called "white goods". Since home appliances are rather expensive and their purchase is associated with risk, households often show brand loyalty and a limited search activity with few alternatives (Bayus, 1992). Home appliances typically have a long lifespan. So, Hennies & Stamminger (2016, p. 76) estimate an average lifespan of a washing machine of 12 years. The decision to buy a home appliance is typically a household decision (Santhi Salomi & Revthi, 2014) and is – to some degree – also influenced by friends and relatives.

Recently, smart home technology is becoming increasingly important for home appliances. So, e.g., voice control of kitchen appliances and virtual lists on smartphones showing which products are currently in the refrigerator is an important trend as well as the ability to track energy consumption (NIQ, 2023). Home appliances are therefore becoming more complex and user-friendly due to technological improvements, particularly with the invention of newer technologies such as the Internet of Things (Aydin & Seker, 2020). In Germany, the trend toward higher-quality, multifunctional appliances is particularly strong, including refrigerators, washer-dryers, stoves and ovens with pyrolytic self-cleaning, and steam-operated appliances (NIQ, 2019). Furthermore, capacity plays an increasingly important role, with increasing demand for larger-capacity appliances. This is intended to simplify and make life more convenient (NIQ, 2023). Thanks to the new insight provided by the tracking option, consumers are increasingly opting for energy-efficient and smart products to save costs in the long term. So, e.g., the share of energy efficiency class A for washing machines, refrigerators, freezers, and dishwashers rose by 14 % to 23 % in 2023. For washing machines, the share of energy efficiency class A was even 64 % (NIQ, 2023). Energy labels are also becoming an increasingly decisive factor. In a GfK study, 82 % of consumers stated that energy labels influence their purchasing decisions; for 70 % of respondents, energy efficiency is even the decisive purchasing criterion. Almost 50 % pay attention to the labels to protect the environment, and 60 % to save on electricity costs (NIQ, 2019).

In order to deal with these current trends, especially the increasing demand for larger-capacity home appliances, a major European producer plans to introduce a larger-capacity refrigerator with a freezer compartment in a form that is relatively unknown in Germany, the so-called multi-door refrigerator. These refrigerators have two doors for a large refrigerator compartment and come in two variants: The cross-shape variant has two additional doors for the freezer compartment, while the French-door variant has two drawers instead. On a rational level, these refrigerators are particularly large, well-structured, and offer good food preservation. On an emotional level, these refrigerators represent freedom, luxury, and individuality. The French-door variant is the classic American model and thus embodies greater individuality and freedom, while the cross-shape variant is particularly structured with its four doors and division into several compartments in the freezer. Some product design aspects are to be cleared before introduction. So, e.g., it is unclear which attribute-levels with respect to accessories, water and ice, and door opening are preferred by the potential customers. Here, a qualitative preliminary study followed by a CBC study should clarify these issues.

As a preliminary study, interviews were conducted with n=38 potential customers, aged 25 to 55 years old. About a quarter of them already owned a multi-door refrigerator. A flexible storage box and a so-called IceCenter were highlighted as important design features. Younger customers preferred a lever handle for door opening, elder customers the recessed grip. The IceCenter was important especially for the younger customers, the accessories for the elder ones. For reasons of confidentiality, the results cannot be discussed in further detail at this point. A quantitative CBC study should clarify whether the results from this preliminary study can be generalized to the target group for this refrigerator.

4.2. Data collection from a human sample

For preparing the main study, a meeting was held with the product manager and the head of the innovation department of the household appliance producer to determine the target group and content of the survey. The target segment was defined by the company as Germans, 25 to 55 years old that spend more than average for home appliances. For creating a CBC questionnaire, Lighthouse Studio 9.15.4 of Sawtooth Software was used (see www.sawtoothsoftware.com). Advantages of this software include the ability to present stimuli visually, web-based data collection, and easy handling. The questionnaire consisted of five parts, beginning with an introduction explaining the structure and objectives of the study. The concept of larger-capacity refrigerators was explained, the so-called multi-door refrigerators, as well as its two possible forms, the cross-shape variant with two additional doors for the freezer compartment and the French-door variant with two drawers instead. Figure 1 summarizes this introduction as given in part 1 of the questionnaire.

This introduction was followed in part 2 by general questions about age and gender to control quotas if needed. Part 3 contained descriptions of the attribute-levels, followed by part 4 with the CBC choice tasks. In the qualitative study, more than 15 different attributes and levels were discussed, but not all of them could be integrated into the CBC task. Table 3 presents the three attributes used in this survey including descriptions of the three levels for each attribute as used in the questionnaire.

The CBC part of the questionnaire consisted of ten choice tasks where respondents were asked to select a preferred option for a multi-door refrigerator (an attribute-level combination based on Table 3) among four presented ones, Figure 2 shows a sample selection for a respondent.

Four options for each choice task were defined in order to force the respondents to make more trade-offs. Furthermore, the questionnaire was only intended to address design preferences, so price was deliberately not included. The non-choice option was not included to avoid its disadvantages, such as avoiding difficult decisions, which leads to a deterioration in validity. Furthermore, the non-choice option does not provide any information about preferences for the attributes of the choice alternatives, which is the main reason for conducting the CBC. Therefore, even if the selected stimulus may not be the best possible overall, an explicit choice should be made from the available sets. The balanced overlap design method was used. This design method makes it possible to present identical attribute-levels within a single choice task – which is, of course, a necessity when there are three levels per attribute and four options to be displayed – while simultaneously aiming to ensure that attribute-level pairs appear with equal frequency across all choice tasks. In addition, the characteristics were shown equally frequently (level balance). 300 different versions were created and the attributes were displayed randomly to re-

You are being asked about a specific type of refrigerator with a freezer compartment, a so-called multi-door appliance, specifically a premium multi-door appliance. This is a particularly spacious and luxurious, and therefore more expensive, refrigerator with a freezer compartment, which is particularly common in America. These appliances are also becoming more popular in Germany and, in a sense, serve as a status symbol, simplifying life with their large capacity and numerous features.

To avoid confusion due to different terms, this refers to French-door and cross-door appliances, i.e., all appliances with two doors at the top of the refrigerator compartment and either two drawers or two additional doors at the bottom for the freezer compartment. The difference from a standard refrigerator-freezer combination is that these appliances have two doors at the top for the refrigerator compartment instead of one, and two drawers or doors in the freezer compartment as well.

Fig. 1: Introduction to multi-door refrigerators and task for respondents as used at the beginning of the CBC study.

Attribute	Levels	Description with visualization	
Accessories	Butter dish	The butter dish can be conveniently transported and opened using the lid with one hand, is shatterproof and dishwasher safe.	
	Cold pack	This serves as an additional cooling reserve in the event of a potential power outage to maintain the cooling quality; it is mounted in the top of the freezer compartment to save space.	
	Flexible storage box	This provides an overview and creates order and is intended for smaller food items, packages, tubes and jars.	
Water and ice	Water dispenser	The water dispenser provides you with fresh, cold water at the outside door.	
	AutoFill	A jug in the refrigerator regularly fills with water, allowing a large amount of cold water to be drawn quickly.	
	IceCenter	In addition to a water dispenser, the IceCenter also includes an ice cube dispenser to produce ice cubes at the touch of a button; for this purpose, part of the volume inside is replaced.	
Door opening	Recessed grip	The door is opened manually; there is no handle; you reach into the recess.	
	Lever handle	The door is opened manually using a handle made of high-quality material.	
	AutoDoor	Automatic opening occurs through knocking or voice command (pairing with an app or voice assistant, such as Alexa, is possible, for example).	

Tab. 3: Attributes and levels for the CBC study.

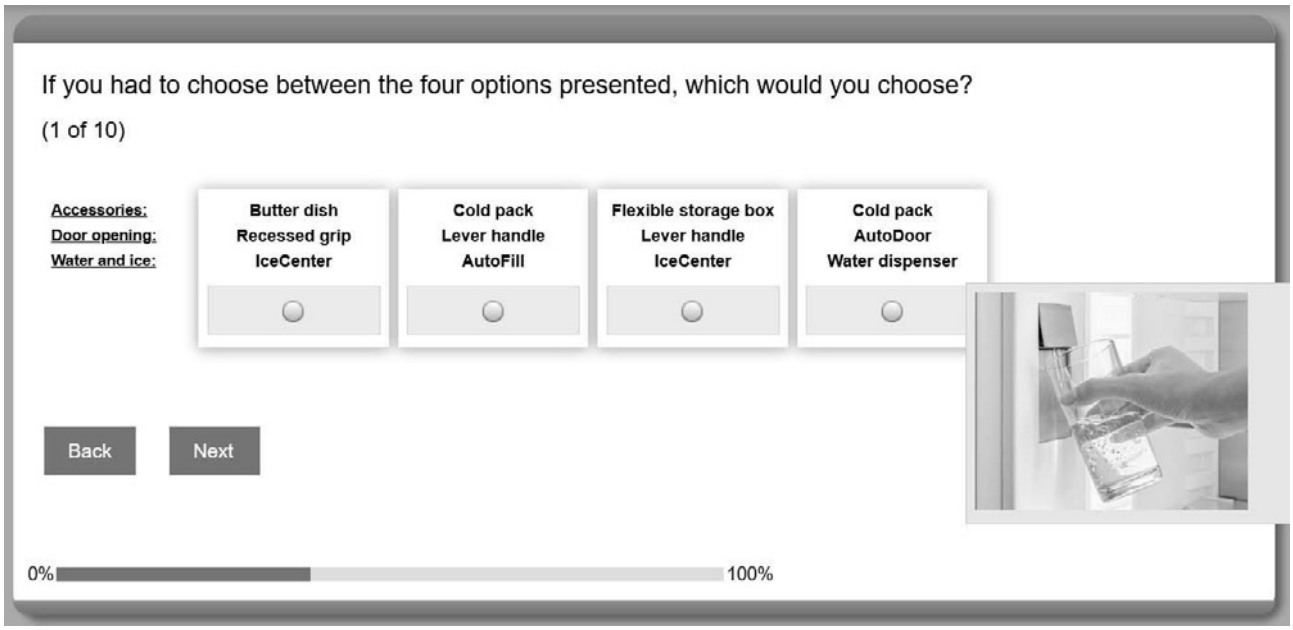


Fig. 2: Sample choice task for a respondent during the survey using Sawtooth Software's CBC module. Additional explanations for the attribute-levels (besides the introductory texts to all attributes and levels) are available for the respondents by so-called "mouse-overs" (here: a visualization is shown for the attribute-level Water dispenser).

duce order effects (Sawtooth Software, 2017). Finally, in part 5, consumer innovativeness, household size, yearly net income, and educational level were surveyed. Consumer innovativeness was measured according to the literature (Klein et al., 2020; Tellis et al., 2009): Respondents had to indicate their agreement with three statements ("I appreciate having new products", "I look forward to trying new products", "I like being exposed to new ideas") on a 5-point rating scale (1 = "strongly disagree", 2 = "disagree", 3 = "neutral", 4 = "agree", 5 = "strongly agree"). The answers could be converted by averaging to measurements on a 5-point consumer innovativeness rating scale (1="low", 2="rather low", 3="medium", 4="rather high", 5="high"). Before the survey started, the questionnaire was successfully tested with 10 respondents of different ages to improve the questionnaire and check for misunderstandings, but no necessary modifications were found.

Data collection took place in summer 2024, via the commercial online access panel Cint (https://de.cint.com/access). The sample should consist of about 600 respondents, 200 aged 25 to 35 years old, 200 aged 36 to 45 years old, and 200 aged 46 to 55 years old. The sample size follows the recommendations by Sawtooth Software (Orme 2020, p. 65), requiring a minimum of n=200 respondents per subgroup. For gender, no quota was defined as proposed by the company decision-makers for being not relevant (based on past studies). During data collection, screeners paid special attention – as discussed by Arndt et al. (2022) – to "speeders," i.e., individuals who completed the survey significantly faster than the majority. These individuals want to complete the survey as quickly as possible, for example, in online access panels, receive their rewards, which results in a low response quality. There are no established standards in the literature for identifying these "speeders," but the median or mean is usually used (Arndt et al., 2022). This was 5:22 minutes for the completers. Individuals who answered the questionnaire more than twice as fast were considered "speeders" and eliminated. As a final control, the times (in seconds) the participants needed for each of the ten CBC tasks were examined to exclude additional respondents who made their selection decisions purely by chance. After deleting n=127 respondents according to the discussed criteria, the panel was reopened, to obtain 600 usable data sets. Table 4 discusses the sample characteristics of the final n=617 respondents, later used for partworth estimation.

4.3. Data collection using GPT-5.2

To make our results from the human and a silicon sample comparable, we use – without the parts where sociodemographic and psychographic characteristics were collected – the same questionnaire and the same sample description (according to Table 2) as prompt for GPT-5.2 as it was used in the online access panel. It should be noted that, in an earlier version of our investigation, we utilized the then-current GPT version (GPT-4o) before switching – as described here – to GPT-5.2. However, the results revealed no significant differences between the two versions. Since GPT-5.2 is not able to respond to a Sawtooth Software online questionnaire directly, we used R code (www.r-project.com) and R packages httr and jsonlite to generate prompts for each respondent in the human sample and to collect the corresponding answers by GPT-5.2. The prompt contained – for each respondent according to the human sample – her/his sociodemographic and psychographic characteristics and the

Characteristic	Levels	Frequencies (Percentages)	Characteristic	Levels	Frequencies (Percentages)
Age	25 to 35 years	200 (32.4%)	Gender	Female	325 (52.7%)
	36 to 45 years	201 (32.6%)		Male	290 (47.0%)
	46 to 55 years	216 (35.0%)		Divers	2 (0.3%)
Monthly gross household income	<1,500 €	73 (11.8%)	Household size	1	120 (19.4%)
	1,500-2,999€	128 (20.7%)		2	179 (29.0%)
	3,000-4,499€	182 (29.5%)		3	134 (21.7%)
	4,500-5.999€	134 (21.7%)		4	141 (22.9%)
	6,000-7,499€	62 (10.0%)		5	32 (5.2%)
	>7,499€	38 (6.2%)		6+	11 (1.8%)
Highest education level	Primary school	2 (0.3%)	Innovativeness (three items according to Tellis et al. 2009, Cronbach's α=0.871)	Low	12 (1,9%)
	Secondary sch.	51 (8.3%)		Rather low	35 (5.7%)
	Med. maturity	199 (32.3%)		Medium	185 (30.0%)
	High school	181 (29.3%)		Rather high	272 (44.1%)
	University	184 (29.8%)		High	113 (18.3%)

Tab. 4: Sample characteristics of the CBC study (n=617).

choice tasks shown during her/his answering of the CBC questionnaire (available through the CHO-files when using Sawtooth Software Lighthouse for data collection). It should be mentioned that when using an online access panel, these sociodemographic and psychographic characteristics of a sample of respondents are available in advance (in order to define a target segment) and that the choice tasks for a respondent are also available in advance (since they stem from a balanced overlap design generated when compiling and hosting a CBC questionnaire, in our case with 300 different choice task sets).

The R program generated for each of the n=617 respondents a corresponding prompt that started with informing GPT-5.2 that he/she is a respondent with defined characteristics and should participate in a survey where preferences for multi-door refrigerators are investigated (see Appendix A for a sample prompt). Similar to the panel questionnaire, the rest of the prompt consists of two additional parts: first, the possible attributes and levels according to Table 1 are explained, then, the chatbot has to act as the described respondent and to select repeatedly (10 times) a preferred option for a multi-door refrigerator of four presented ones. Afterwards, the prompt was passed to GPT-5.2 using the ChatGPT API and the answers were stored like the answers from human respondents for further analysis. To prevent the data collection from GPT ordering effects, the presentation of the four stimuli in the 10 choice tasks was randomized. This proved to be necessary since without this reordering of the stimuli in each choice task, the first stimulus in the choice task would have been selected only with a probability of 11 % to 14 % compared to its expected selection probability of 25 %. In all GPT-5.2 calls the temperature was set to 1.0 as in other applications (Arora et al., 2025; Brand et al., 2023) for generating silicon samples. This is a medium value for the output softmax functions. Small values (e.g., 0.25) are expected to generate conservative answers with few variations, large values generate creative answers with high variations. We tested alternative values for the temperature parameter (from 0.25 to 1.5) but found only few differences in GPT-5.2's answers so we decided to use the value 1.0 across all calls.

The prompt for generating answers to the choice tasks follows the usual and proven successful chain-of-thought strategy (see the review by Yu et al., 2023) that is based on human step-by-step thinking and integrates intermediate steps to guide the selection: "Please go through these 10 tasks and answer the following questions for each: Compare the Accessories attribute for the options shown. Which would you prefer? Compare the Water and ice attribute for the options shown. Which would you prefer? Compare the Door Opening attribute for the options shown. Which would you prefer? If you had to choose between the four options presented in each of these 10 tasks, which would you choose?" (see Appendix A for a complete sample prompt). GPT-5.2's answers to the prompt could be given – for controlling the reasoning process – in a long version (were each selection of an option in a choice task was explained, referring to attribute-levels and the respondent's characteristics) or short version (giving only the 10 numbers of the selected option in each choice task, taking into account the reordering in each choice task, easier to handle for the follow-up activities). A closer look at the answers demonstrated that GPT-5.2 was able to justify its selections based on the attribute-levels and the respondent's characteristics.

As an experimental factor, the answers by GPT-5.2 were generated under varying amounts of information provided since we assume from Arora et al. (2025)'s and Brand et al. (2023)'s investigations that this can have a major influence on the quality of results:

- In the "without any additional information" case, GPT-5.2 only receives the already discussed respon-

dent's characteristics (see the GPT-5.2 prompt in Appendix A).

- In the "with market-releated information from company experts" case, we summarized in a short addition to the prompt the information that was available to the manufacturer before the preliminary and main study (see Section 4.1): So, e.g., the company's experts assumed from other surveys that the Accessories attribute is the least important for customers, and the Water and ice attribute is the most important. They also assumed that the Flexible storage box and the IceCenter levels are popular among all respondents and the Cold pack and AutoDoor levels are particularly popular among customers that demonstrate a high innovativeness. It is explicitly mentioned that these are assumptions by the company's experts (see the GPT-5.2 prompt in Appendix B).
- In the "with information from a preliminary study" case, the results from the preliminary study were made available: The respondents liked the Flexible storage box, the Ice center, and the Water dispenser levels. The younger respondents also the Lever handle, the elder the Recessed handle and the Lever handle levels (see the GPT-5.2 prompt in Appendix C).

A complete run of the R program (with the generation of the prompts for 617 respondents, the GPT-5.2 call for each respondent, and the processing and storing of answers in a CHO-file) took about 34 minutes. The GPT-5.2 API within a run consumed about 860,000 ChatGPT API tokens during one complete run (available for about 0.20 US$ costs).

Besides the collection of choices from the human sample as the ground truth case and the generation of choices for the three GPT-5.2 cases, as a fifth case, also random choices were generated with R for later comparisons. The answers of all five cases were stored by the R program in separate text files in Sawtooth Software's CHO-format that allows to analyze the collected or generated choices later using Sawtooth Software's CBC/HB system (Sawtooth Software, 2021).

5. Results

5.1. Comparison of partworth estimates derived from the human and the silicon samples

Sawtooth Software's CBC/HB system (Sawtooth Software, 2021) was used to estimate partworths at the individual level from the human sample (choices from the online panel), the silicon sample with three different amounts of information provided (choices "without any additional information", "with market-releated information from company experts", and "with information from a preliminary study") as well as from random choices. In all cases, we used the answers to eight of the ten choice tasks to estimate the individual partworths by hierarchical Bayes estimation (leaving out the choice tasks 3 and 8). The two holdout choice tasks can then be used to control the predictive validity of the estimated partworths at the individual level by counting the so-called first choice hits (FCHs which indicates the share of identical predicted and observed choices). Moreover, the so-called RLH (root likelihood) indicates for each respondent as a measure of internal validity the probability the respondent would have made the selections he/she made, given their preference, or "utility" scores based on the choice tasks used for estimating the partworths. A random score for four options in a choice task would be 1/4 = 0.25. Table 5 summarizes the mean attribute-level partworth estimates, the attribute importances, and the fit criteria derived from these choices.

The results for the human sample (choices from the online panel) indicate an acceptable internal validity average RLH score of 0.522 (standard deviation 0.161). The FCHs for choice task 3 and for choice task 8 (63.0 %) indicate a good predictive validity. Most preferred attribute-levels (on average) are the Flexible storage box and the IceCenter. To a lesser extent, also the Lever handle, the Water dispenser, and the Recessed grip levels are preferred. Compared to results from random choices, the information that is gained from the CBC study is enormous: The CBC/HB system provides a partworth estimation with a very low internal validity average RLH score of 0.279 (standard deviation 0.034) and low FCHs (21.6 %). The partworth estimates for each attribute are close together and the importance of the attributes is more or less equal (34.8 %, 32.9 %, 32.4 %). The correlation between the partworths estimated from the online panel and from the random choices is very low (0.173). Moreover, the partworth estimates are also very low, which indicates in a multinomial logit model that predicted choice probabilities among stimuli/product are equally distributed.

The estimates from the silicon samples also show interesting results. In the "without any additional information" resp. the "with market-related information from company experts" cases, the average RLH scores are slightly better than the one from random choices (mean 0.392 with a standard deviation of 0.098 resp. mean 0.497 with a standard deviation of 0.119). But the first choice tasks still need improvement (52.8 % resp. 20.3 %). The partworths show a preference for the Lever handle, the Water dispenser, and the Butter dish levels, that partly was also expected from the past studies and the preliminary study, but the partworths only show a low correlation (0.144 resp. 0.651) to the "ground truth" from the online access panel, in the first use case even lower when compared to partworths derived from random choices (see also Figure 3 for this comparison, for perfect correspondence, attribute-levels should lie on the bisector of the first and third quadrants). The average partworth estimates demonstrate rather high values, an aspect that reflects that GPT-5.2 made its selections in the choice tasks with an assumption of preferred attribute-levels, however these assumptions (derived from the

Attribute	Levels and importance, fit criteria	Random choices	GPT choices in the "without any additional information" case	GPT choices in the "with market-related information from company experts" case	GPT choices in the "with information from a preliminary study" case	Choices from the online panel
Acces-sories	Butter dish	.010(.147)	.051(.061)	-.258(.032)	-.221(.039)	-.110(.131)
	Cold pack	-.017(.160)	-.017(.071)	.124(.043)	-.044(.050)	.008(.145)
	Flexible storage box	.007(.174)	-.034(.058)	.134(.034)	.265(.030)	.102(.146)
	Importance	.348(.145)	.156(.069)	.415(.047)	.486(.048)	.346(.164)
Water and ice	Water dispenser	.007(.154)	.133(.065)	-.048(.041)	.032(.036)	.022(.135)
	AutoFill	-.008(.153)	-.081(.053)	-.123(.037)	-.154(.024)	-.086(.121)
	IceCenter	.001(.154)	-.052(.074)	.171(.025)	.122(.030)	.064(.175)
	Importance	.329(.142)	.254(.076)	.298(.043)	.278(.038)	.328(.164)
Door opening	Recessed grip	.029(.154)	-.097(.166)	-.148(.023)	-.061(.038)	.020(.123)
	Lever handle	-.012(.150)	.297(.062)	.011(.032)	.139(.028)	.024(.132)
	AutoDoor	-.018(.151)	-.200(.152)	.138(.029)	-.078(.036)	-.044(.200)
	Importance	.324(.149)	.589(.079)	.286(.040)	.236(.042)	.327(.165)
Internal validity: RLH value		.279(.034)	.392(.098)	.497(.119)	.509(.127)	.522(.161)
Internal validity; First choice hits		21.6%	52.8%	20.3%	68.0%	63.0%
External validity: Pearson's correlation with online-panel partworth estimates at the individual level		-.009(.394)	.079(.402)	.258(.366)	.369(.366)	-
External validity: Pearson's correlation with online-panel partworth estimates at the aggregate level		.173	.144	.651*	.928***	-
External validity: Spearman's correlation with online-panel partworth estimates at the individual level		-.012(.387)	.059(.386)	.238(.367)	.365(.354)	-
External validity: Spearman's correlation with online-panel partworth estimates at the aggregate level		.167	.183	.533*	.933***	-

Tab. 5: Mean attribute-level partworths (standard deviations) and attribute importances (standard deviations) estimated from different choice generations/collections as well as fit criteria (standard deviations) for internal and external validity (: significant at the 0.05 level, ***: significant at the .001 level).*

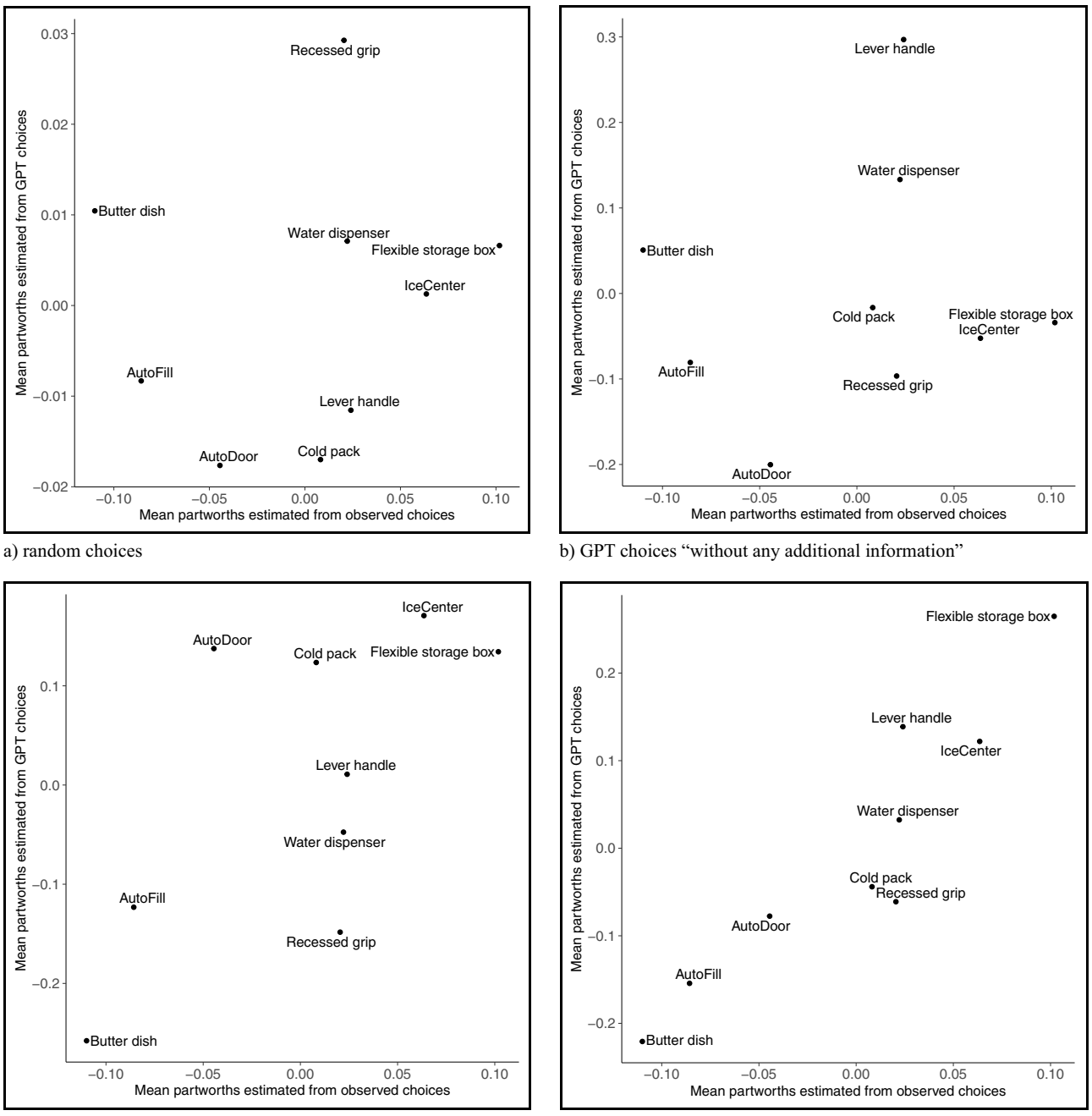
LLM) obviously were not compatible with the ground truth.

In the "with information from a preliminary study" case, the fit criteria and the estimated partworths come much closer to the ground truth, at least at the aggregate level (correlation 0.928, significant at the 0.001 level). Obviously was GPT-5.2 able to draw useful conclusions from the presented information that goes beyond the LLM.

However, in all cases, the correlation between the estimates from the silicon sample choices and the estimates from the online panel choices are rather low (from 0.079 to 0.369 on average) which indicates a low reproduction of "the ground truth" at the individual level: Just providing sociodemographic and some psychographic data to GPT-5.2 is not enough to produce digital twins with an external validity at the individual level.

5.2. Methodological implications

Our methodological comparison clearly shows that – even with an up-to-date and interactively improved chain-of-thought prompting strategy – GPT-5.2 is not able to generate valid choices in a home appliance CBC study if "market-related information from company experts" or "from a preliminary study" is missing. In this case, besides the trained LLM, respondent characteristics like age, gender, household size, income, and innovativeness are the sole contexts available, consequently, the estimated partworths resemble more estimates from randomly generated choices than estimates from "real" choices (our so-called ground truth), collected from respondents in an online access panel with paid participants. This is especially true when we hope that the silicon sample contains reproductions of respondents at the individual level: Sociographic and few psychographic information is not enough for GPT-5.2 to generate similar preferences and choices at the individual level.



a) random choices

b) GPT choices "without any additional information"

c) GPT "with market-related information from experts"

d) GPT choices "with inf. from a preliminary study"

Fig. 3: Mean partworths estimated from observed online panel choices (x-axis) versus mean partworths estimated (y-axis) from a) random choices, b) GPT choices in the "without any additional information" case, c) GPT choices in the "with market-related information from company experts" case, d) GPT choices in the "with information from a preliminary study" case.

However, if "market-related information from company experts" or "from a preliminary study" is available, the resemblance increases considerably, at least at the aggregate level, showing a highly significant Pearson correlation coefficient of .928 between the partworth estimates from observed and GPT-5.2 generated choices. These findings support Brand et al. (2023)'s, Goli & Singh (2024)'s, and Arora et al. (2025)'s findings with respect to silicon samples but – for the first time – the silicon sample was developed as a sample of digital twins the person-by-person replicates the human sample in a CBC study (using balanced overlap designs and similar numbers of choice tasks). Even though this failure at the individual level is obvious, our findings at the aggregate level are in contrast to Sarstedt et al. (2024)'s and Park et al. (2024)'s findings who "cast doubts on the validity of using LLMs as a general replacement for human participants in the social sciences," but used older LLMs for their comparisons.

Also, it should be mentioned (as discussed in the sample application) that still some problems have to tackled when using LLMs for the generation of answers to choice tasks in a CBC study: ChatGPT is still very sensible to the presentation of information in the prompts:

"First" and "last" options in a choice task are selected more often than other options. Or, ChatGPT answers in a free format to choice tasks, even if asked just to answer with the number of the selected option requiring additional text processing of the answer.

5.3. Managerial implications

At a first glance, our comparison has beneficial managerial implications: Somehow expensive CBC studies can be replaced by "market-related information from company experts" and (cheaper) "preliminary studies" (e.g. focus groups, pre-tests) without loosing the access to valid individual level partworth estimates for large samples of respondents which form an excellent basis for further managerial decision support, market simulation and product line design as usual in traditional CBC studies (Green & Krieger, 1991; Kohli & Sukumar, 1990; Orme, 2020): Based on estimated partworths at the individual level, descriptions of the status quo market (current offers in a market, described in terms of attribute-levels), and cost information, derived market share, market volume, and profit predictions are possible and can be used to design new products that maximize these predictions.

6. Conclusions and outlook

Our investigation generated for the first time – to the best of our knowledge – silicon samples that simulate human choice decisions in a real-world CBC study (using Sawtooth Software for data collection and analysis). Brand et al. (2024) and other authors only used greatly simplified choice tasks in their analysis.

At the aggregate level (mean and groupwise partworths across the sample), the estimation based on the human sample leads to similar results as the estimation based on the silicon sample. This is especially true, when "market-related information from company experts" or "from preliminary studies" is made available. One can argue (see, e.g., Sarstedt et al. 2024) that silicon sample should enrich a human sample, not replace.

The similarity depends also on the data with which GenAI was trained: For standard markets, the estimation should work, but not for very specific products and target segments not represented in the training data (see, e.g., Brand et al. 2024).

At the individual level (partworths for each replicated respondent), the results are disappointing: GenAI produced consistent selections across the choice tasks (documented in the RLH value for internal validity) but the correlations between individual partworth estimates based on human and GPT-5.2 generated choices remained low. The limited data that describe the synthetic respondent seems to be the reason: If more past answers to questions were available (preliminary information at the individual level), better results could be possible.

One should test whether elaborated digital twins of online access panel members (with answers from past surveys) are superior. However, the General Data Protection Regulation defines clear limits.

Appendix A

Sample ChatGPT prompt in the "without information from past and preliminary studies" case

---

You are Person 1 and are German, male, 46–55 years old.

Two people live in your household.
Your highest level of education is a university entrance qualification/high school diploma/master's degree or equivalent.
You earn €4,500-€5,999 gross per month.
Your innovativeness is rather high.

You are being asked about a specific type of refrigerator with a freezer compartment, a so-called multi-door appliance, specifically a premium multi-door appliance. This is a particularly spacious and luxurious, and therefore more expensive, refrigerator with a freezer compartment, which is particularly common in America. These appliances are also becoming more popular in Germany and, in a sense, serve as a status symbol, simplifying life with their large capacity and numerous features.

To avoid confusion due to different terms, this refers to French-door and cross-door appliances, i.e., all appliances with two doors at the top of the refrigerator compartment and either two drawers or two additional doors at the bottom for the freezer compartment. The difference from a standard refrigerator-freezer combination is that these appliances have two doors at the top for the refrigerator compartment instead of one, and two drawers or doors in the freezer compartment as well.

The goal of this survey is to find out which of the attributes listed below are particularly important to you.

The survey consists of two sections. In the first section, the attributes and levels will be explained to you. In the second section you will then be asked to make several choices.

First Section

The following explains the possible attributes and levels of the multi-door appliance, on which we would like to hear your opinion. This involves a total of three attributes, each with three different levels:

1) The attribute Accessories has the levels: a) Butter dish, b) Cold pack, c) Flexible storage box.
2) The attribute Water and ice has the levels: a) Water dispenser, b) AutoFill, c) IceCenter.
3) The attribute Door opening has the levels: a) Recessed handle, b) Lever handle, c) AutoDoor.

Please read the following information carefully and memorize their respective attributes and levels!

Attribute 1) Accessories:
Level a) Butter dish: Thanks to its one-handed operation, the butter dish can be conveniently transported by the lid and opened easily. It is shatterproof and dishwasher-safe.
Level b) Cold pack: This serves as an additional cold reserve in the event of a potential power outage to maintain the cold quality; it is mounted on the top of the freezer compartment to save space.
Level c) Flexible storage box: This provides a clear overview and creates order and is designed for smaller food items, packages, tubes, and glasses.

Attribute 2) Water and ice
Level a) Water dispenser: The water dispenser provides fresh, cold water at the outside door.
Level b) AutoFill: A pitcher in the refrigerator regularly fills with water, allowing a larger amount of cold water to be quickly dispensed.
Level c) IceCenter: In addition to a water dispenser, the IceCenter also includes an ice cube dispenser for ice cubes at the touch of a button; for this purpose, part of the volume inside is replaced.

Attribute 3) Door opening
Level a) Recessed handle: The door is opened manually; there is no handle; you reach into the recess.
Level b) Lever handle: The door is opened manually using a handle made of high-quality material.
Level c) AutoDoor: Automatic opening occurs by knocking or voice command (pairing with an app or voice assistant, such as Alexa, is possible, for example).

Second Section

You are considering a premium multi-door purchase. The following question suggests four refrigerator options, each with variations of the potential attribute-levels just explained.

Please select the refrigerator option you would most likely choose from the 10 tasks (1 to 10) below.

Task 1:
Option 1: Butter dish; IceCenter; AutoDoor
Option 2: Flexible storage box; IceCenter; Lever handle
Option 3: Cold pack; AutoFill; AutoDoor
Option 4: Butter dish; Water dispenser; Recessed handle

Task 2:
Option 1: Cold pack; IceCenter; Recessed handle
Option 2: Flexible storage box; Water dispenser; AutoDoor
Option 3: Butter dish; AutoFill; Lever handle
Option 4: Cold pack; Water dispenser; Recessed handle

Task 3:
Option 1: Flexible storage box; AutoFill; AutoDoor
Option 2: Butter dish; IceCenter; AutoDoor
Option 3: Butter dish; Water dispenser; Lever handle
Option 4: Cold pack; AutoFill; AutoDoor

Task 4:
Option 1: Butter dish; IceCenter; Lever handle
Option 2: Cold pack; Water dispenser; Recessed handle
Option 3: Flexible storage box; AutoFill; Lever handle
Option 4: Flexible storage box; IceCenter; Recessed handle

Task 5:
Option 1: Cold pack; IceCenter; AutoDoor
Option 2: Butter dish; Water dispenser; Recessed handle
Option 3: Flexible storage box; AutoFill; AutoDoor
Option 4: Cold pack; AutoFill; Lever handle

Task 6:
Option 1: Flexible storage box; Water dispenser; AutoDoor
Option 2: Butter dish; IceCenter; Recessed handle
Option 3: Flexible storage box; AutoFill; Lever handle
Option 4: Cold pack; AutoFill; Recessed handle

Task 7:
Option 1: Flexible storage box; AutoFill; Recessed handle
Option 2: Flexible storage box; Water dispenser; Lever handle
Option 3: Cold pack; Water dispenser; AutoDoor
Option 4: Butter dish; Water dispenser; AutoDoor

Task 8:
Option 1: Cold pack; AutoFill; Recessed handle
Option 2: Butter dish; IceCenter; Lever handle
Option 3: Cold pack; Water dispenser; AutoDoor
Option 4: Butter dish; IceCenter; Recessed handle

Task 9:
Option 1: Flexible storage box; IceCenter; Recessed handle
Option 2: Cold pack; IceCenter; Lever handle
Option 3: Butter dish; Water dispenser; AutoDoor
Option 4: Flexible storage box; AutoFill; AutoDoor

Task 10:
Option 1: Cold pack; IceCenter; AutoDoor
Option 2: Flexible storage box; Water dispenser; Recessed handle
Option 3: Flexible storage box; AutoFill; Recessed handle
Option 4: Butter dish; IceCenter; Lever handle

Which attribute would be most important to you, and which would be second most important?
Which level of the Accessories attribute would you prefer?
Which level of the Water and ice attribute would you prefer?
Which level of the Door opening attribute would you prefer?

Please go through these 10 tasks and answer the following questions for each:
Compare the Accessories attribute for the options shown. Which would you prefer?
Compare the Water and ice attribute for the options shown. Which would you prefer?
Compare the Door opening attribute for the options shown. Which would you prefer?

If you had to choose between the four options presented in each of these 10 tasks, which would you choose?

Please answer in the following format for all tasks without variation: Task number, Option number

Appendix B

ChatGPT prompt in the „with information from past studies" case (different start into the second section)

...

Second Section

You are considering a premium multi-door purchase. The following question suggests four refrigerator options, each with variations of the potential attribute-levels just explained.

Expert interviews revealed that the Accessories attribute is the least important for customers, and the Water and ice attribute is the most important. It is also assumed that the Flexible storage box level is the most popular for the Accessories attribute, and the IceCenter level is the most popular for the Water and ice attribute. The Cold pack and AutoDoor levels are particularly popular with customers with a high propensity for innovation.

Please select the refrigerator option you would most likely choose from the 10 tasks below.

Task 1
...

Appendix C

ChatGPT prompt in the „with information from a preliminary study" case (different start into the second section)

...

Second Section

You are considering a premium multi-door purchase. The following question suggests four refrigerator options, each with variations of the potential attribute-levels just explained.

From a preliminary survey, we know the following:

The 25 to 35 year-old age group particularly likes the Flexible storage box level for the Accessories attribute, the IceCenter and Water

dispenser for Water and ice, and the Lever handle for Door opening. The Water and ice attribute is most important to this age group.

The 36 to 45 year-old age group particularly likes the Flexible storage box for the Accessories attribute, the IceCenter and Water dispenser for Water and ice, and the Lever handle for Door opening. The Accessories attribute is most important to this age group.

The 46 to 55 year-old age group particularly likes the Flexible storage box and the Cold pack levels for the Accessories attribute, the IceCenter and Water dispenser for Water and ice, and the Recessed handle and Lever handle for Door opening. The Accessories attribute is most important to this age group.

Please select the refrigerator option you would most likely choose from the 10 tasks below.

Task 1:

...

References

Allenby, G. M., Arora, N., & Ginter, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, *32*(2), 152–162.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Arndt, A.D., Ford, J. B., Babin, B. J., & Luong, V. (2022). Collecting samples from online services: How to use screeners to improve data quality. *International Journal of Research in Marketing*, *39*(1), 117–133.

Arora, N., Chakraborty, I., & Nishimura, Y. (2025). AI-human hybrids for marketing research: Leveraging Large Language Models (LLMs) as collaborators. *Journal of Marketing*, *89*(2), 43–70.

Aydin, N., & Seker, S. (2020). Assessing customer preferences on a new design of refrigerator using conjoint analysis. In C. Kahraman & S. Cebi (Eds.), *Customer oriented product design – Intelligent and fuzzy techniques* (pp. 417–428). Cham, Switzerland: Springer.

Baier, D., & Brusch, M. (2021). Conjointanalyse: Erfassung von Kundenpräferenzen im Überblick. In D. Baier & M. Brusch (Hrsg.), *Conjointanalyse: Methoden-Anwendungen-Praxisbeispiele* (pp. 3–34). Berlin, Heidelberg, Deutschland: Springer.

Bayus, B. L. (1992). Brand loyalty and marketing strategy: An application to home appliances. *Marketing Science*, *11*(1), 21–38.

Bouschery, S. G., Blazevic, V., & Piller, F. T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management*, *40*(2), 139–153.

Brand, J., Israeli, A., & Ngwe, D. (2023). Using LLMs for market research. *Harvard Business School Working Paper*, *23–062*.

Callegaro, M., Baker, R., Bethlehem, J., Göritz, Anja, S., Krosnick, J. A., & Lavrakas, P. J. (2014). Online panel research: History, concepts, applications and a look into the future. In M. Callegaro, R. Baker, J. G. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Wiley series in survey methodology. Online panel research: A data quality perspective* (pp. 1–22). Hoboken, NJ: John Wiley & Sons Inc.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology, 15(3), 39:1–39:45.*

Filippi, S. (2023). Measuring the impact of ChatGPT on fostering concept generation in innovative product design. *Electronics*, *12*(16), 3535.

Galarraga, I., Heres, D. R., & Gonzalez-Eguino, M. (2011). Price premium for high-efficiency refrigerators and calculation of price-elasticities for close-substitutes: a methodology using hedonic pricing and demand systems. *Journal of Cleaner Production*, *19*(17–18), 2075–2081.

Goli, A., & Singh, A. (2024). Frontiers: Can large language models capture human preferences? *Marketing Science*, *43*(4), 709–722.

Göritz, A. S. (2004). Recruitment for online access panels. *International Journal of Market Research*, *46*(4), 411–425.

Göritz, A. S. (2006). Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science*, *1*(1), 58–70.

Green, P. E., & Krieger, A.M. (1991). Product design strategies for target-market positioning. *Journal of Product Innovation Management*, *8*(3), 189–202.

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, *8*(3), 355–363.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, *5*(2), 103–123.

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, *54*(4), 3–19.

Green, P. E., & Wind, Y. (1975). New ways to measure consumer judgements. *Harvard Business Review*, *53*, 107–117.

Grewal, D., Satornino, C. B., Davenport, T., & Guha, A. (2025). How generative AI is shaping the future of marketing. *Journal of the Academy of Marketing Science*, *53*(3), 702–722.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y.(2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint, arXiv:2301.07597*.

Hennies, L., & Stamminger, R. (2016). An empirical survey on the obsolescence of appliances in German households. *Resources, Conservation and Recycling*, *112*, 73–82.

Hensel-Börner, S., & Sattler, H. (2000). Validity of customized and adaptive hybrid conjoint analysis. In R. Decker & W. Gaul (Eds.), *Classification and Information Processing at the Turn of the Millenium* (pp. 320–329), Berlin, Heidelberg: Springer.

Herring, D., Robertson, M., Gatzer, S., & La Motte, H. de. (2025). *What's trending in white goods? Durable, efficient, and affordable*. Annual McKinsey Global State of White Goods Survey 2025, McKinsey & Company, retrieved from www.mckinsey.com/industries/industrials-and-electronics/our-insights/whats-trending-in-white-goods-durable-efficient-and-affordable.

Jeong, N., & Lee, J. (2024). An aspect-based review analysis using ChatGPT for the exploration of hotel service failures. *Sustainability*, *16*(4), 1640.

Kaiser, C., Kaiser, J., Manewitsch, V., Rau, L., & Schallner, R. (2025). Simulating human opinions with large language models. In UMAP (Ed.), *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 82–86). New York, NY, USA: ACM.

Kato, T., & Miura, T. (2021). The impact of questionnaire length on the accuracy rate of online surveys. *Journal of Marketing Analytics*, *9*(2), 83–98.

Klein, F. F., Emberger-Klein, A., & Menrad, K. (2020). Indicators of consumers' preferences for bio-based apparel: A German case study with a functional rain jacket made of bioplastic. *Sustainability*, *12*(2), 675.

Koc, E., Hatipoglu, S., Kivrak, O., Celik, C., & Koc, K. (2023). Houston, we have a problem! The use of ChatGPT in responding to customer complaints. *Technology in Society*, *74*, 102333.

Kohli, R., & Sukumar, R. (1990). Heuristics for product-line design using conjoint analysis. *Management Science*, *36*(12), 1464–1478.

Kulshreshtha, K., Bajpai, N., Tripathi, V., & Sharma, G. (2019). Consumer preference for eco-friendly appliances in trade-off: a conjoint analysis approach. *International Journal of Product Development*, *23*(2/3), 212–243.

Louviere, J. J., & Woodworth, G. (1983). Design and analysis of

simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research*, *20*(4), 350–367.

McKinsey & Company. (2023). *The economic potential of generative AI: The next productivity frontier.* Retrieved from https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#/.

NIQ. (2019). *Haushaltsgroßgeräte: 3 Trends treiben den Markt*. Retrieved from nielseniq.com/global/de/news-center/2019/haushaltsgrosgerate-3-trends-treiben-den-markt/.

NIQ. (2023). *Nachhaltige und smarte Haushaltsgeräte liegen im Trend*. Retrieved from nielseniq.com/global/de/news-center/2023/nachhaltige-und-smarte-haushaltsgerate-liegen-im-trend/.

Orme, B. K. (2020). *Getting started with conjoint analysis: Strategies for product design and pricing research* (4th ed.). Manhattan Beach, CA: Research Publishers LLC.

Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, *56*(6), 5754–5770.

Rao, V. R. (2026). *Applied Conjoint Analysis*. Second Edition, Cham, Switzerland: Springer.

Razali, M. A. S., Kamaludin, M., & Azlina, A.A. (2022). Consumer preference for energy label in the purchase decision of refrigerator: A discrete choice experiment approach in the east coast, Malaysia. *International Journal of Energy Economics and Policy*, *12*(3), 441–450.

Roberts, J. H., Kayande, U., & Stremersch, S. (2014). From academic research to marketing practice: Exploring the marketing science value chain. *International Journal of Research in Marketing*, *31*(2), 127–140.

Rudolph, L., & Seelkopf, L. (2025). Maximize power: Online survey panels and the price-quality trade-off. *SocArXiv*, *13.02.2025*.

Santhi Salomi, R., & Revthi, B. (2014). Customer's buying behavior towards home appliances: An empirical study. *International Journal of Management Research*, *2*(2), 301–328.

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, *41*(6), 1254–1270.

Sawtooth Software (2017). *The CBC System for Choice-Based Conjoint Analysis – Technical Paper Version 9*. Orem, UT: Sawtooth Software.

Sawtooth Software (2021). *The CBC/HB System – Technical Paper Version 5.6*. Provo, UT: Sawtooth Software.

Sawtooth Software (2025). *Report on Panel Sample Usage and Conjoint Analysis Usage 2025*. Retrieved from drive.google.com/file/d/10hwG7J3cn4dc2KFKUlfQrPqLMCif5Ri5/view.

Selka, S., & Baier, D. (2014). Kommerzielle Anwendung auswahlbasierter Verfahren der Conjointanalyse: Eine empirische Untersuchung zur Validitätsentwicklung. *Marketing ZFP – Journal of Research and Management*, *36*(1), 54–66.

Steiner, M., & Meißner, M. (2018). A user's guide to the galaxy of conjoint analysis and compositional preference measurement. *Marketing ZFP – Journal of Research and Management*, *40*(2), 3–25.

Sundberg, L., & Holmström, J. (2024). Innovating by prompting: How to facilitate innovation in the age of generative AI. *Business Horizons*, *67*(5), 561–570.

Tafesse, W., & Wood, B. (2024). Hey ChatGPT: an examination of ChatGPT prompts in marketing. *Journal of Marketing Analytics*, *12*(4), 790–805.

Tellis, G. J., Yin, E., & Bell, S. (2009). Global consumer innovativeness: Cross-country differences and demographic commonalities. *Journal of International Marketing*, *17*(2), 1–22.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30. 31st Annual Conference on Neural Information Processing Systems (NIPS 2017): Long Beach, California, USA, 4–9 December 2017*. Red Hook, NY: Curran Associates Inc., 1–11.

Viglia, G., Adler, S. J., Miltgen, C. L., & Sarstedt, M. (2024). The use of synthetic data in tourism. *Annals of Tourism Research*, *108*, 103819.

Wang, A., Morgenstern, J., & Dickerson, J. P. (2024). Large language models that replace human participants can harmfully misportray and flatten identity groups. *arXiv*, *2402.01908.*

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences of the United States of America*, *122*(47), e2518075122.

Yu, Z., He, L., Wu, Z., Dai, X., & Chen, J. (2023). Towards better chain-of-thought prompting strategies: A survey. *arXiv preprint, arXiv:2310.04959.*

Zha, D., Yang, G., Wang, W., Wang, Q., & Zhou, D. (2020). Appliance energy labels and consumer heterogeneity: A latent class approach based on a discrete choice experiment in China. *Energy Economics*, *90*, 104839.

Keywords

Choice-Based Conjoint Analysis (CBC); Generative Artificial Intelligence (GenAI); Large Language Model (LLM); Generative Pretrained Transformer (GPT); silicon samples; home appliances